doi:10.3969/j.issn.1001-0505.2010.05.003

基于语音信号与心电信号的多模态情感识别

黄程韦[1] 金 赟[1,2] 王青云[1] 赵 力[1] 邹采荣[1]

([1]东南大学水声信号处理教育部重点实验室,南京 210096)
([2]徐州师范大学物理与电子工程学院,徐州 221116)

摘要:通过采集与分析语音信号和心电信号,研究了相应的情感特征与融合算法.首先,通过噪声刺激和观看影视片段的方式分别诱发烦躁情感和喜悦情感,并采集了相应情感状态下的语音信号和心电信号.然后,提取韵律、音质特征和心率变异性特征分别作为语音信号和心电信号的情感特征.最后,利用加权融合和特征空间变换的方法分别对判决层和特征层进行融合,并比较了这2种融合算法在语音信号与心电信号融合情感识别中的性能.实验结果表明:在相同测试条件下,基于心电信号和基于语音信号的单模态情感分类器获得的平均识别率分别为71%和80%;通过特征层融合,多模态分类器的识别率则达到90%以上;特征层融合算法的平均识别率高于判决层融合算法.因此,依据语音信号、心电信号等不同来源的情感特征可以构建出可靠的情感识别系统.

关键词:情感识别;多模态;判决层融合;特征层融合
中图分类号:TP391.4 **文献标志码**:A **文章编号**:1001-0505(2010)05-0895-06

Multimodal emotion recognition based on speech and ECG signals

Huang Chengwei[1] Jin Yun[1,2] Wang Qingyun[1] Zhao Li[1] Zou Cairong[1]

([1]Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing 210096, China)
([2]School of Physics and Electronics Engineering, Xuzhou Normal University, Xuzhou 221116, China)

Abstract: Through collecting and analyzing speech signals and electrocardiography (ECG) signals, emotion features and fusion algorithms are studied. First, annoyance is induced by noise stimulation and happiness is induced by comedy movie clips. The corresponding speech signals and ECG signals are recorded. Then, prosodic features and voice quality features are adopted for speech emotional features, and heart rate variability (HRV) features are used for ECG emotional features. Finally, the decision level fusion and the feature level fusion are accomplished by the weighted fusion method and the feature transformation method, respectively. The performances of the two fusion methods in speech emotion and ECG emotion recognition are compared. The experimental results show that for the same testing set, the average recognition rates of the single modal classifier based on the ECG signals and the single modal classifier based on the speech signals reach 71% and 80%, respectively, while that of the multi-modal classifier with the feature level fusion of the speech signals and the ECG signals achieves above 90%. The average recognition rate of the feature level fusion algorithm is higher than that of the decision level fusion algorithm. The different signal channels such as speech signals and ECG signals show a promising improvement in building a reliable emotion recognition system.

Key words: emotion recognition; multimodal; decision level fusion; feature level fusion

收稿日期:2010-01-26. **作者简介**:黄程韦(1984—),男,博士生;赵力(联系人),男,博士,教授,博士生导师,zhaoli@seu.edu.cn.
基金项目:国家自然科学基金资助项目(60472058,60975017)、江苏省自然科学基金资助项目(BK2008291).
引文格式:黄程韦,金赟,王青云,等.基于语音信号与心电信号的多模态情感识别[J].东南大学学报:自然科学版,2010,40(5):895-900.
[doi:10.3969/j.issn.1001-0505.2010.05.003]

情感的自动识别是实现自然人机交互的关键技术之一[1]. 目前,情感识别向多模态方向发展,单一的依靠表情、语音或者生理参数来进行情感识别的研究取得了一定的成果,但是如何将这些不同性质的情感信号融合,达到信息上的互补,从而建立一个鲁棒性强、识别率高的系统还需要进一步深入研究.

利用人脸表情、语音、眼动、姿态和生理信号等多个通道的情感信息之间的互补性来提高分类器的识别性能,能够提高分类器的识别率;此外,在实际噪声环境下,当某一个通道的特征受到干扰或缺失时,这种方法还能使分类器具有良好的鲁棒性. Hoch 等[2]通过融合语音与表情信息,在车载环境下进行了正面(愉快)、负面(愤怒)与平静等3种情感状态的识别. Busso 等[3]分析了单一的语音情感识别与人脸表情识别在识别性能上的互补性,分别通过特征层融合与决策层融合进行基于多模态信息的情感识别. Wagner 等[4]通过融合肌动电流、心电、皮肤电阻和呼吸4个通道的生理参数,获得了92%的融合识别率.

本文以语音信号与心电信号(CEG)为基础,对烦躁、喜悦和平静3种情感状态进行识别,研究了心电信号与语音信号的融合情感识别及相应的融合算法和情感特征. 通过在特征层面和判决层面进行融合,比较了基于语音信号与心电信号2种单模态分类器的识别率及其之间的互补性,并建立了基于多模态信息的分类器,以提高情感识别性能. 这种基于多模态信息的分类器在实际应用中具有重要意义. 例如,在噪声等环境干扰下,当语音信号的采集受到影响时,生理信号为情感识别提供了重要的依据. 此外,目前基于心电信号等生理参数的情感识别能分辨的情感种类较少,识别率相对较低,与语音特征融合后,可使识别性能得到较大提高.

1 情感诱发与数据采集

高自然度的情感数据采集是目前情感识别领域中受到重点关注的问题之一,越来越多的研究者通过诱发的方式来采集情感数据. 本文中,通过让被试人员在噪声环境下进行四则运算来诱发烦躁情感,通过观看喜剧片段诱发喜悦情感,并通过充分休息采集平静状态下的数据. 实验流程如图1所示. 参与实验的被试为5名男性和5名女性,年龄为20~40岁,健康状况良好,近期无药物服用. 实验中要求被试人员读出指定的文本语句,录制烦躁、平静和愉快3种情感状态下的语音数据. 在实验全过程中记录心电数据,并截取每条语音数据开始前30 s到结束后30 s时间段内的心电数据,与相对应的语音数据绑定存储. 由于情绪一般持续1~2 min,而HRV频谱等心电特征的提取一般需要至少1 min的数据,因此在实验中截取1~2 min的心电数据作为一条样本.

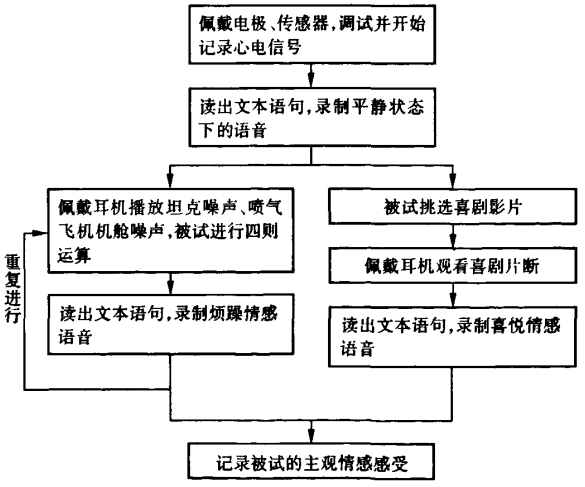

图1 情感数据采集流程

2 情感特征提取

情感特征的优劣以及情感特征提取是否全面,直接影响到情感识别的性能. 本文从语音信号与心电信号2个方面提取并构造了用于识别烦躁、喜悦与平静状态的特征. 基于语音信号的情感识别研究相对较多,基音、能量、共振峰以及语速等参数是受到广泛认同的有效的语音情感特征. 除了这些基本

的语音情感特征外，还提取了谱能量分布、声门波、谐波噪声比（HNR）等方面的音质特征参数，用于加强对效价维度的区分能力（烦躁与喜悦在效价维度上差异较大）. 目前，通过心电信号来进行情感识别的研究还较为缺乏，常用的心电情感特征有心率异常性（HRV）方面的时域、频域特征，更多有效的心电情感特征有待发掘，心电情感特征中的年龄差异等因素也有待研究. 本文除了提取常见的 HRV 特征外，还提取了心电信号的若干混沌特征，用于进行烦躁、喜悦与平静等 3 种情感的研究.

2.1　语音情感特征

以往对情感特征参数的有效提取主要以韵律特征为主，然而近年来通过深入研究发现，音质特征和韵律特征相互结合才能更准确地识别情感[5]. Tato 等[6]研究发现，音质类特征对于区分激活维接近的情感有较好的效果，证实了共振峰等音质类特征与效价维度的相关性较强. 本文使用了 74 个全局统计特征，其中前 36 个特征为韵律特征，后 38 个特征为音质特征.

特征 1 ~ 10：短时能量及其差分的均值、最大值、最小值、中值、方差.

特征 11 ~ 25：基音及其一阶、二阶差分的均值、最大值、最小值、中值、方差.

特征 26：基音范围.

特征 27 ~ 36：发音帧数、不发音帧数、不发音帧数和发音帧数之比、发音帧数和总帧数之比、发音区域数、不发音区域数、发音区域数和不发音区域数之比、发音区域数和总区域数之比、最长发音区域数、最长不发音区域数.

特征 37 ~ 66：第 1，2，3 共振峰及其一阶差分的均值、最大值、最小值、中值、方差.

特征 67 ~ 69：250 Hz 以下的谱能量百分比、650 Hz 以下的谱能量百分比及 4 kHz 以上的谱能量百分比.

特征 70 ~ 74：谐波噪声比的均值、最大值、最小值、中值、方差.

语音特征参数的提取算法可以参照文献[7]. 虽然 650 ~ 4 000 Hz 频段涉及第 1 共振峰和几乎全部的第 2 共振峰，但是其能量受到文本内容变化的影响较大，且主要随着音位信息的变化而变化[8]，因此在构造频谱能量的分频段特征时未采用该频段内的能量百分比，

采用了 4 kHz 以上频谱能量特征是因为根据 Pittam 等的研究结果[9]可知，这一部分频段能量的增加能反映激励程度的提高，可用于区分悲伤与愤怒等.

谐波噪声比以往常用于诊断喉部疾病，是衡量说话人嗓音沙哑程度的一个特征. Biemans[10]将谐波噪声比作为音质特征，用于评价语音的音质.

2.2　心电情感特征

大多内脏器官都是受交感神经和副交感神经双重支配的，心电信号也不例外. 心脏的每次跳动都是由窦房结的起搏引起的，窦房结内起搏细胞固有的节律性受自主神经调节，交感神经增快其自发激动，副交感神经减慢激动. 研究表明，情绪的变化对心电信号有一定影响. 心率变异性等指标被越来越多的研究者用于情绪的生理心理学研究中.

在 HRV 的频域分析中，短时 HRV 功率谱的高频成分（HF）是与呼吸同步的，可定量估计呼吸性心律失常，代表副交感神经活动指数，并可作为监测心脏迷走神经活动水平的定量指标；低频成分（LF）代表了交感神经活动的指数，随交感神经活动的增强而增加. 低频/高频能量比则可作为评价心脏迷走-交感神经均衡性的定量指标，在一定程度上反映出情感状态的变化. 3 种情感状态下 HRV 的低频/高频能量比如图 2 所示.

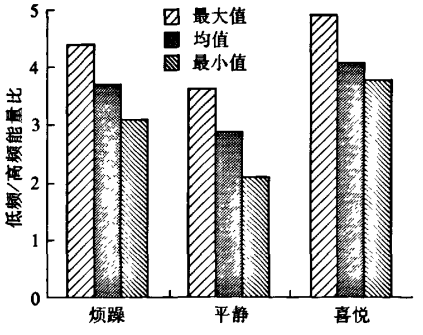

图 2　3 种情感状态下的 HRV 特征分析

本文中除了采用常见的 HRV 波、R 波、T 波等心电特征参数进行情感状态的分析外，还研究了心电的混沌特征，提取了心电的关联维数作为反映情绪变化的生理指标之一. ECG 信号不是一个单纯的周期信号，其非线性研究主要集中在非线性动力学参数的计算上，如分形维数、Lyapunov 指数等.

这里采用关联维数来描述心电信号的混沌特征. 关联维特征是从单变量时间序列中提取维数信息，表示系统在多维空间中的疏密程度，反映了系统中点与点之间的关联程度. 实际计算中心电信号的嵌入维数设定为 8 较为合理，采用 Grassberger-Procaccia 算法（G-P 算法）得到的 3 种情感状态下的关联维数如表 1 所示.

表 1 心电信号的关联维特征分析

关联维特征	烦躁	平静	喜悦
均值	2.553	2.875	2.621
最大值	3.041	3.433	3.142
最小值	2.163	2.359	2.310

对烦躁、喜悦和平静 3 种情感状态的识别,本文采用了以下 23 个心电特征参数.

特征 1 ~8:关联维数及 Lyapunov 指数的最大值、最小值、均值、方差.

特征 9 ~12:RR 间期的最大值、最小值、均值、方差.

特征 13 ~15:HRV 的低频能量、高频能量、低频/高频能量比.

特征 16 ~23:T 波及 R 波能量的最大值、最小值、均值、方差.

心电参数的提取算法可以参照文献[11].

3 多模态融合识别算法

为充分利用基于语音信号与心电信号的情感特征,现分别通过判决层融合算法和特征层融合算法来进行语音与心电的双模态数据的融合识别.

3.1 判决层融合算法

在判决层融合算法中,首先分别设计出基于语音信号的情感分类器和基于心电信号的情感分类器,将 2 个分类器依据一定的准则进行判决融合,得到最终的识别结果(见图 3).

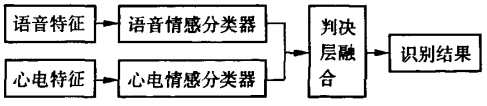

图 3 判决层融合算法

本文中待识别的情感类别包括烦躁、喜悦和平静 3 个类别. 对于 2 种分类器,均采用高斯混合模型(Gaussian mixture model,GMM)来进行每种情感类别的概率模型训练. 高斯混合模型是 M 个成员密度的加权和,可以用如下形式表示:

$$P(\boldsymbol{X}_t \mid \lambda) = \sum_{i=1}^{M} a_i b_i(\boldsymbol{X}_t) \tag{1}$$

式中,$\boldsymbol{X}_t$ 为第 t 个 D 维随机向量;$b_i(\boldsymbol{X}_t)$ $(i=1,2,\cdots,M)$ 为成员密度;a_i 为混合权值,且 $\sum_{i=1}^{M} a_i = 1$. 每个成员密度均为一 D 维变量的关于均值矢量 $\boldsymbol{U}_i$ 和协方差矩阵 $\boldsymbol{\Sigma}_i$ 的高斯函数,可表示为

$$b_i(\boldsymbol{X}_t) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{X}_t - \boldsymbol{U}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{X}_t - \boldsymbol{U}_i) \right\} \tag{2}$$

完整的高斯混合密度由所有成员密度的均值矢量、协方差矩阵和混合权值参数化而成. 这些参数聚集在一起可表示为

$$\lambda_i = \{a_i, \boldsymbol{U}_i, \boldsymbol{\Sigma}_i\} \qquad i = 1,2,\cdots,M \tag{3}$$

GMM 模型的参数估计采用 EM 算法迭代计算获得.

当存在噪声干扰时,语音分类器的性能会发生下降;当心电电极受到抖动、碰撞或者仪器内部的基线漂移干扰时,心电分类器的性能会发生下降. 这就需要在选择判决层融合算法时,考虑评价各个子分类器在某一时刻的置信度,并根据分类器的输出置信度来进行融合判决. 这里采用一种样本自适应的方法来衡量分类器对当前样本的判决是否可靠,对置信度高的分类器给予较高的融合权值,对于置信度低的分类器赋予较低的融合权值[12]. 子分类器(语音分类器与心电分类器)给出的 3 种情感类别的 GMM 似然度分别记为 $P(\boldsymbol{X} \mid \lambda_k)$,其中 $k=1,2,3$ 时分别对应了这 3 种情感类别. 当属于各个类别的 GMM 似然度基本相等或差别不大时,认为该样本很可能处于概率分布模型的重叠区域,该子分类器的判决置信度较低;当分类器给出的似然度值较为分散时,则认为样本处于概率分布模型的非重叠区域,该子分类器的判决置信度较高. 因此,每个子分类器的融合权值可表示为

$$w_j = \frac{\sum_{1 \le m < n \le 3} \left| \ln(P(\boldsymbol{X} \mid \lambda_m)) - \ln(P(\boldsymbol{X} \mid \lambda_n)) \right|}{\left| \sum_{k=1}^{3} \ln(P(\boldsymbol{X} \mid \lambda_k)) \right|} \tag{4}$$

式中,j 为子分类器编号,$j=1,2$.

当分类器判决越可靠时,差值越大;反之当差值越小时,说明样本距离重叠区域越近,分类可靠性越差. 定义了子分类器的融合权值后,对每个子分类器的判决进行加权融合,则最终的分类器融合判决输出为

$$i^* = \arg\max \left\{ \sum_{j=1}^{2} w_j P^j(\boldsymbol{X} \mid \lambda_k) \right\} \tag{5}$$

3.2 特征层融合算法

在特征层融合算法中,并不设计多个单模态的情感分类器,而是将来自多个通道的大量情感特征通过特征选择算法进行优化选取,使用单个分类器对由语音数据与心电数据共同构成的最佳特征组进行分类识别(见图 4). 对双模态数据组成的特征,采用高斯混合模型进行训练与识别.

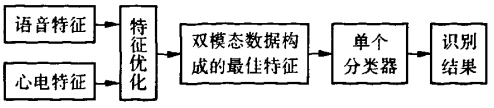

图 4　特征层融合算法

用于情感识别的原始语音特征包括韵律特征和音质特征等 74 维；与情感状态有关的心电特征包括非线性特征、时域特征、频域特征等 23 维. 特征层融合的关键是对这些原始的情感特征进行优化选取，使得语音和心电双模态数据高效组合，提高特征与情感的相关程度. 进行特征优化的方法通常有 2 种：包装法（wrapper）和滤波法（filter）. 其中，包装法与情感识别系统后端所采用的识别器相关性较大，采用不同的模式识别方法，会对特征选择的结果产生较大的影响[13]；滤波法能够在一定的准则下找出最佳意义上的情感特征，因而其通用性较好. 本文采用 PCA 方法进行语音特征与心电特征的融合与降维. Ververidis 等[14]研究发现，4 ~ 5 个特征已经足以描述情感的分布，并可获得较好的识别率. 因此，本文截取 PCA 变换的前 5 个维度来构成识别用的特征矢量. 在 PCA 变换后的前 3 个维度构成的特征空间中，3 类心电情感样本的分布如图 5 所示.

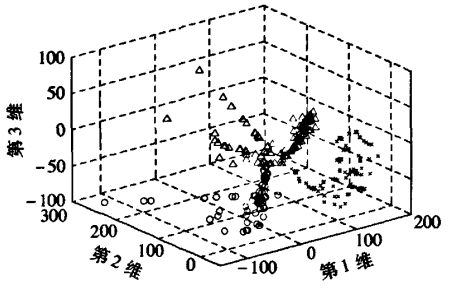

图 5　语音与心电双模数据的特征层融合

4　实验与结果

实验中采用 GMM 模型来拟合各类情感的概率分布结构，GMM 混合度设为 6，即高斯分量的个数为 6. 训练样本集和测试样本集均包含 3 种类别：烦躁样本集、平静样本集和喜悦样本集，分别用 Ⅰ，Ⅱ，Ⅲ 表示. 每个训练样本集包含每种情感的 300 条语音样本与 300 条心电样本，每个测试样本集包含每种情感的 100 条语音样本与 100 条心电样本. 实验中采用相同的训练集与测试集对单模态、多模态分类器进行测试，结果见表 2 和表 3.

表 2　基于语音信号的单模态分类器的识别率　%

样本集	烦躁	平静	喜悦
Ⅰ	81	9	10
Ⅱ	13	76	11
Ⅲ	7	11	82

表 3　基于心电信号的单模态分类器的识别率　%

样本集	烦躁	平静	喜悦
Ⅰ	71	15	14
Ⅱ	19	69	12
Ⅲ	11	15	74

由表 2 可知，基于语音信号的单模态分类器的平均识别率达到 80%；烦躁情感在实际中具有重要的应用价值，其平均识别率为 81%，说明本文中采用的语音情感特征与烦躁情感的相关性较高，能够用于烦躁情感的识别. 由表 3 可知，基于心电信号的单模态分类器对 3 种情感状态的区分能力较弱，平均识别率略高于 71%，因此单纯依靠心电数据的情感识别在实际应用中会遇到一定的困难，需要与其他类型的情感数据相结合来进行多模态的情感识别，以提高识别率与可靠性.

在多模态的情感识别中，采用图 3 与图 4 所示的识别系统，通过判决层融合与特征层融合 2 种算法来进行情感识别. 实验结果显示，相比单模态分类器，多模态分类器的识别性能有了显著提高. 其中，判决层融合算法的平均识别率达到 88%，基于特征层融合的平均识别率则达到 90% 以上. 虽然基于心电信号的单模态分类器对情感的识别能力有限，但是心电信号提供了一部分语音数据所不能替代的生理信息. 通过加入心电数据后情感识别系统的性能得到了明显提高，相比传统的语音单模态识别系统，其平均识别率提高约 9%.

2 种融合算法的识别结果见表 4 和表 5. 可以看出，判决层融合算法在平均识别率上略低于特征层融合算法，后者对喜悦状态的识别率高达 94%. 对于烦躁状态的识别，2 种融合算法的性能均较高，说明多模态融合算法获得了预期的效果. 判决层融合算法的优势在于，每个分类器都是相互独立的，当某一通道的情感数据无法获取或质量较低时，判决层仍然能够进行情感识别，鲁棒性较高. 特征层融合算法能够在一定条件下获得最佳的特征压缩与优化性能，在识别测试中识别率略高于判决层融合算法.

表4 判决层融合算法的识别率 %

样本集	烦躁	平静	喜悦
Ⅰ	88	10	2
Ⅱ	5	87	8
Ⅲ	8	2	90

表5 特征层融合算法的识别率 %

样本集	烦躁	平静	喜悦
Ⅰ	90	6	4
Ⅱ	5	88	7
Ⅲ	4	2	94

5 结语

本文通过诱发手段采集了语音与心电的情感数据,建立了多模态的情感识别系统,并对语音情感特征和心电情感特征进行了分析与优化选择.以往的情感识别研究大多依靠单通道的数据来进行,如表情识别、语音情感识别等.本文通过将多种不同性质和不同来源的情感特征进行融合,实现多模态识别,从而提高系统的识别率和鲁棒性.将语音信号与心电信号融合,建立情感识别系统,能够有效地利用这2种不同的生物信号进行情感识别.实验结果表明,融合后系统的识别率得到了明显提升.判决层融合算法和特征层融合算法在实验中显示出不同的识别特性,其中特征层融合算法的平均识别率较高.何种语音特征和心电特征能够有效地反映出烦躁等具有实际意义的情感,还需要进一步深入研究;心电情感特征与情感维度(如激活维、效价维、控制维等)之间的关系也是今后值得探讨的问题.

参考文献 (References)

[1] Zeng Z, Pantic M, Roisman G I, et al. A survey of affect recognition methods: audio, visual and spontaneous expressions [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(1): 39-58.

[2] Hoch S, Althoff F, McGlaun A, et al. Bimodal fusion of emotional data in an automotive environment [C]// *Proceedings of the* 2005 *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Philadelphia, Pennsylvania, USA, 2005: 1085-1088.

[3] Busso C, Deng Z, Yildirim S, et al. Analysis of emotion recognition using facial expressions, speech and multimodal information [C]//*Proceedings of the Sixth International Conference on Multimodal Interfaces*. Pennsylvania, USA, 2004: 205-211.

[4] Wagner J, Kim J, Andre E. From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification [C]// *Proceedings of the* 2005 *IEEE International Conference on Multimedia & Expo*. Amsterdam, the Netherlands, 2005: 940-943.

[5] Khiet T. How does real affect affect affect recognition in speech? [D]. Enschede, the Netherlands: Center for Telematics and Information Technology of University of Twente, 2009.

[6] Tato R, Santos R, Kompe R, et al. Emotion space improves emotion recognition [C]//*Proceedings of the* 2002 *International Conference on Speech and Language Processing*. Denver, Colorado, USA, 2002: 2029-2032.

[7] 赵力. 语音信号处理[M]. 北京:机械工业出版社, 2003.

[8] Schuller B, Rigoll G, Lang M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture [C]//*Proceedings of the* 2004 *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Montreal, Canada, 2004: 577-580.

[9] Pittam J, Scherer K R. *Vocal expression and communication of emotion* [M]. New York, USA: Guilford Press, 1993: 185-198.

[10] Biemans M. Gender variation in voice quality [D]. Nijmegen, the Netherlands: Department of Linguistics of Radboud University Nijmegen, 2000.

[11] 景慎旗. 基于LabVIEW的多生理信号采集与处理的研究[D]. 南京:东南大学生物科学与医学工程学院,2009.

[12] 蔡莉莉. 基于数据融合的语音情感分析与识别[D]. 南京:东南大学信息科学与工程学院,2005: 46-48.

[13] Peng Hangchuan, Long Fuhui, Ding Chris. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(8): 1226-1238.

[14] Ververidis D, Kotropoulos C, Pitas I. Automatic emotional speech classification [C]//*Proceedings of the* 2004 *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Montreal, Canada, 2004: 593-596.